Attorney Docket No. 826.1907

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Patent Application of:

Andreas SAVVA

Application No.: Group Art Unit: Not Yet Assigned

Filed: November 26, 2003 Examiner: Not Yet Assigned

For: THE APPARATUS AND THE METHOD FOR INTEGRATING NICS WITH RDMA CAPABILITY BUT NO HARDWARE MEMORY PROTECTION IN A SYSTEM WITOUT DEDICATED MONITORING PROCESSES

SUBMISSION OF CERTIFIED COPY OF PRIOR FOREIGN APPLICATION IN ACCORDANCE WITH THE REQUIREMENTS OF 37 C.F.R. § 1.55

Commissioner for Patents
PO Box 1450
Alexandria, VA 22313-1450

Sir:

In accordance with the provisions of 37 C.F.R. § 1.55, the applicant(s) submit(s) herewith a certified copy of the following foreign application:

Japanese Patent Application No(s). 2002-357449

Filed: December 10, 2002

It is respectfully requested that the applicant(s) be given the benefit of the foreign filing date(s) as evidenced by the certified papers attached hereto, in accordance with the requirements of 35 U.S.C. § 119.

Respectfully submitted,

STAAS & HALSEY LLP

Date: November 26, 2003 By: _____

J. Randall Beckers
Registration No. 30,358

1201 New York Ave, N.W., Suite 700
Washington, D.C. 20005
Telephone: (202) 434-1500
Facsimile: (202) 434-1501

JAPAN PATENT OFFICE

This is to certify that the annexed is a true copy of the following application as filed with this office.

Date of Application: December 10, 2002

Application Number: Patent Application No. 2002-357449
[ST.10/C] [JP2002-357449]

Applicant(s): FUJITSU LIMITED

September 25, 2003

Commissioner,
Japan Patent Office Yasuo IMAI

Certificate No.P2003-3078526

日本国特許庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日
Date of Application: ２００２年１２月１０日

出願番号
Application Number: 特願２００２－３５７４４９

[ST. 10/C]: ［ＪＰ２００２－３５７４４９］

出願人
Applicant(s): 富士通株式会社

２００３年　９月２５日

特許庁長官
Commissioner,
Japan Patent Office 今井康夫

出証番号　出証特２００３－３０７８５２６

【書類名】　特許願

【整理番号】　0252079

【提出日】　平成14年12月10日

【あて先】　特許庁長官殿

【国際特許分類】　H04L 29/08

【発明の名称】　ＲＤＭＡ機能を持ったＮＩＣをハードウェアメモリ保護を行わないで、専用のモニタプロセスなしにシステムに組み込むための装置

【請求項の数】　5

【発明者】

【住所又は居所】　神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内

【氏名】　アンドレアス　サバ

【特許出願人】

【識別番号】　000005223

【氏名又は名称】　富士通株式会社

【代理人】

【識別番号】　100074099

【住所又は居所】　東京都千代田区二番町８番地２０　二番町ビル３Ｆ

【弁理士】

【氏名又は名称】　大菅　義之

【電話番号】　03-3238-0031

【選任した代理人】

【識別番号】　100067987

【住所又は居所】　神奈川県横浜市鶴見区北寺尾７－２５－２８－５０３

【弁理士】

【氏名又は名称】　久木元　彰

【電話番号】　045-573-3683

【手数料の表示】

【予納台帳番号】　012542

【納付金額】　21,000円

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

【物件名】　要約書　1

【包括委任状番号】　9705047

【プルーフの要否】　要

【書類名】　　明細書

【発明の名称】　　ＲＤＭＡ機能を持ったＮＩＣをハードウェアメモリ保護を行わないで、専用のモニタプロセスなしにシステムに組み込むための装置

【特許請求の範囲】

【請求項１】ＲＤＭＡ機能を有した、ホストと複数のホストからなるネットワーク内の該ホストに設けられた装置であって、

該ネットワーク内のホストが起動したとき、該ホストが起動したことを示す第１のメッセージを、該ネットワーク内の全ての該複数のホストに送信する手段と、

該複数のホストからの該ホストへのＲＤＭＡアクセスを利用不可能とする手段と、

該ホストに第２のメッセージを送信することによって、第１のメッセージに応答する手段と、

該ホストが、該複数のホストからのＲＤＭＡアクセスを受付可能であることを示す第３のメッセージを、該複数のホストの全てから第２のメッセージを受信し、ＲＤＭＡ機能を利用可能とした後、該複数のホストの全てに送信する手段と、を備えることを特徴とする装置。

【請求項２】前記装置は、前記ホストのドライバに含まれることを特徴とする請求項１に記載の装置。

【請求項３】他のホストにＲＤＭＡアクセスを行うための情報を有する翻訳保護テーブル手段を更に有し、

前記第１のメッセージを受信したとき、該第１のメッセージを送信したホストに関する情報を該翻訳保護テーブル手段からクリアし、該ホストへのＲＤＭＡアクセスを不可能にすることを特徴とする請求項１に記載の装置。

【請求項４】前記第２のメッセージは、アクノレッジメント、ノンアクノレッジメント、及び、前記複数のホストから送られた前記第１のメッセージの一つであり、ノンアクノレッジメントは、ハードウェアによって生成されることを特徴とする請求項１に記載の装置。

【請求項５】前記ホストは、ＲＤＭＡ機能と他のメッセージ通信機能を有す

るネットワークインターフェースカードを有し、ＲＤＭＡ機能と他のメッセージ通信機能の初期化は独立して行われることを特徴とする請求項１に記載の装置。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、ＲＤＭＡ機能を有するホストマシンをネットワークに組み込み、あるいは、再組み込みするための装置に関する。

【０００２】

【従来の技術】

リモートＤＭＡ（ＲＤＭＡ：Remote Direct Memory Access）は、ネットワークインタフェースカード（ＮＩＣ：Network Interface Card）に対する、特に、ＲＤＭＡ拡張機能を通常のＡＰＩに付加するという最近の提案においては、重要な要求となってきている。ＲＤＭＡをサポートするＮＩＣは、ローカルメモリへのＲＤＭＡアクセスが適切に有効化することを確保するために、メモリプロテクション、例えば、翻訳保護テーブル（ＴＰＴ：translation and protection table）をサポートすることが要求される。このようなテーブルを実装することは、ハードウェアコストを増大し、特に、周辺バス（例えば、ＰＣＩ：Peripheral Component Interconnect）へ接続されるＮＩＣの性能に影響を与える可能性がある。ハードウェアベースのＴＰＴの代替法法としては、他のＮＩＣへのＲＤＭＡアクセスを有効化するための機能を果たすＮＩＣドライバを構成することである。この場合、全体のシステムは、ＲＤＭＡアクセスがホストをクラッシュさせるようには生じないことを確保するために協同しなくてはいけない。

【０００３】

この問題に対する解決方法は、ホスト上で動作するドライバにコピーが保持される、システム全体をカバーする保護テーブルの形で実現されるであろう。ドライバは、テーブルのローカル部を更新し、変更を遠隔の度ラバに反映する。ホストが再起動しない限りは、問題はない。しかし、ホスト、例えば、ＨＯＳＴ＿Ａが再起動された場合には、システム内の残りのホストは、すぐには、このことを認知しない、あるいは、認知できない。ＨＯＳＴ＿Ａが、ＲＤＭＡアクセスを再

起動の直後に可能とすると、保護テーブル内の古い情報（ＨＯＳＴ＿Ａが再起動される前）に基づいたリモートホストから始められたＲＤＭＡ動作は、ＨＯＳＴ＿Ａをクラッシュさせることがある。

【０００４】

例えば、ホスト（ネットワーク内のノード）が再起動されたとき、ホストの内部ＲＤＭＡ設定、すなわち、登録されたメモリ領域は解除され、初期化される。しかし、他のホストのＲＤＭＡ設定が更新されていないと、他のホストは、ＲＤＭＡを使って、最近再起動したホストのメモリにアクセス使用とするかもしれない。このとき、当該ホストのメモリ割当ては、再起動後初期化されており、再起動されたホストのメモリ割当ての初期化が他のホストの設定に反映されていないので、他のホストの再起動したホストのメモリへの、古い設定によって決定されたアドレスへのアクセスは、再起動したホストの動作に重要な情報を上書きしてしまうかもしれず、再起動したホストをクラッシュさせるかもしれない。

【０００５】

従って、再起動されたホストが他のホストによってそのように認識され、再起動したホストに関連するシステムの保護テーブル内の情報が、再起動したホストが完全にシステムに再び組み込まれるまで、無効化されることを確保することが本質的である。

【０００６】

【発明が解決しようとする課題】

通常の解決方法は、おそらくハートビートユーザレベルプロセスと共に、リモートホストの状態の変化をドライバに通知する各ノード上のユーザレベルプロセス（Ｕｎｉｘ（登録商標）におけるデーモンプロセスのような）の形を採用するかもしれない。例えば、システムは、各ホストにＮＩＣ制御プロセスとハートビートプロセスを有する。ＮＩＣ制御プロセスは、ＮＩＣを初期化する機能を持ち、ＮＩＣドライバに、システムにおける例えば、ホストの再起動などの変化を通知する。各ハートビートプロセスは、ハートビートプロセスとメッセージを交換することによって、他のホストの状態を監視し続ける。例えば、ホスト、ＨＯＳＴ

＿Ａが、落ちた場合には、ハートビートメッセージは送られない。所定の時間後、他のハートビートプロセスは、ＨＯＳＴ＿Ａが落ちたことを断定する。各ハートビートプロセスは、ＮＩＣ制御プロセスにそのことを通知し、これは次にドライバに通知して、ＨＯＳＴ＿Ａへの全てのアクセスをブロックする。ＨＯＳＴ＿Ａが復帰すると、そのハートビートプロセスは再びメッセージを送信し始める。他のハートビートプロセスは、メッセージを受け取り、これらのＮＩＣ制御プロセスに、これを通知し、こんどは、ドライバにこれを通知して、ＨＯＳＴ＿Ａへのアクセスを再び可能とする。しかし、このアプローチには幾つかの問題がある。例えば、ＮＩＣの初期化は、ユーザレベルプロセスであるＮＩＣ制御プロセスに依存しているため、ＮＩＣが初期化されるまでに時間がかかる。ホストの負荷が大きい場合には、このプロセスは、ちょうど良く動作するようにはスケジューリングされいないかもれいない（システム応答が遅い）。この間、ＮＩＣは使用できない。また、ＮＩＣ制御プロセスあるいはハートビートプロセスのいずれかが落ちるかもしれない。ＮＩＣ制御プロセスが落ちることは、システムイベントに対するホストの応答を含み、例えば、再起動したホストへのアクセスが長時間再開されないかもしれない。ハートビートプロセスが落ちることは、ホストのクラッシュと間違えられる可能性があり、従って、全体のシステム性能に影響する。すなわち、ユーザレベルプロセスを使用することは、システム全体の信頼性を低くすることになる。

【０００７】

他の問題は、潜在的に、ＲＤＭＡリクエストによって使用されるものとは異なる、組み込みプロトコルのための通信経路を用い、ＮＩＣのＲＤＭＡ機能が利用可能とされる前に、実行途中のＲＤＭＡアクセスが無いことを確保することができないユーザレベルプロセスに起因することになろう。

【０００８】

ノードが再起動されたとする。そのノードのドライバは、以前の状態の全ての情報を失う。特に、全ての、以前に登録したメモリ領域と、他のノードに登録されているメモリ領域に関する全ての情報を失う。しかし、リモートノードは、そのノードが再起動されたことを知らないかもしれない。あるドライバが当該ホス

トが再起動されたことを知り、他のドライバが知らないようなシステムに再起動されたホスト上のＮＩＣを再組み込みする問題は、解決する必要がある。ターゲットであるＮＩＣに対して、動作途中であるＲＤＭＡ動作が無いことを確保しなければならないという意味において、解決方法は安全でなくてはならない。もし、そのようなＲＤＭＡ動作が有ると、これらは、ホストが再起動される前に始動されており、古い情報に基づいており、ホストをクラッシュさせる可能性がある。再組み込みの終わりには、そのＮＩＣドライバは、システムの保護テーブルの最新のコピーを有していなければならず、システムの一部である、他の全てのホストは、そのＮＩＣがシステムの一部として機能していることを知らなければらない。この問題をユーザレベルモニタプロセスで解決することは、安全な時間になるまで、ＮＩＣの全体の初期化を送らせることになる。ユーザレベルプロセスが通信するためには、各ホストは、他のＮＩＣと、ＲＤＭＡ機能なしに、適合されていなければならず、全体のハードウェアコストを増加させる。

【０００９】

本発明の課題は、ＲＤＭＡ機能を有するＮＩＣを持つホストを、ＮＩＣの初期化を制御するのに、ユーザレベルプロセスに頼らないで、ネットワークに安全に組み込むための装置を提供することである。

【００１０】

【課題を解決するための手段】

本発明の装置は、ＲＤＭＡ機能を有している複数のホストからなるネットワーク内のホストにある装置であって、ネットワーク内のホストが起動されたとき、ネットワーク内の複数のホストの全てに、当該ホストが起動されたことを示す代１のメッセージを送信する手段と、複数のホストから当該ホストへのＲＤＭＡアクセスを全て利用不可能にする手段と、当該ホストに第２のメッセージを送ることによって、第１のメッセージに応答する手段と、当該ホストが複数のホストからのＲＤＭＡアクセスを受け付ける準備ができたことを示す第３のメッセージを、複数のホストの全てから第２のメッセージを受け取り、ＲＤＭＡ機能が利用可能とされた後に、全ての複数のホストに送信する手段とを備えることを特徴とする。

【００１１】

本発明によれば、ネットワーク内のホストが起動される毎に、ホストは、ネットワークの全てのホストにこれを通知する。従って、全てのホストは、どのホストが起動されたかを知ることができ、当該ホストへのＲＤＭＡアクセスに使用される情報を、正しくクリアし、更新することができる。結果として、起動されたホストは、ネットワークに安全に組み込み、あるいは、再組み込みされるが、クラッシュは生じない。起動された、あるいは、再起動されたホストのネットワークへの組み込み、あるいは、再組み込みは、同様に行うことができる。

【００１２】

このアプローチの独自性の一部は、ＳＥＮＤ／ＲＥＣＥＩＶＥ機能と、ＲＤＭＡ機能の初期化を分離したことにある。ＳＥＮＤ／ＲＥＣＥＩＶＥ初期化は、ＮＩＣ自身がそのためのプロトコルを使って使用されるように実行され、従って、全体のハードウェアコストを下げる。ＲＤＭＡ機能を有しない他のＮＩＣが当該プロトコルを実行する必要はない。更に、ＮＩＣ自体がこのプロトコルを実行するのに使われるため、プロトコルが終了すれば、システム内にＲＤＭＡアクセスが処理されずに残っていることがないことの保証を、ユーザプロセスが使用されうる場合よりも簡単に達成することができる。更に、このプロトコルが依然処理中であっても、ＳＥＮＤ／ＲＥＣＥＩＶＥを介して、通常の通信をＮＩＣを持ち手開始することが可能である。また、ユーザプロセスに対する依存性を減少させることで、システム内の部品の数を減らすことができ、全体の信頼度を上げることができる。この機能をカーネルに含ませることにより、ドライバの開発コストは上がるが、他のコストは、プロトコルが簡単であるため、かなり小さくなる。

【００１３】

【発明の実施の形態】

ここで、再起動されたホストのＮＩＣをシステムに安全に再組み込みし、これらのＲＤＭＡ機能を利用可能にすることを確保するためにユーザレベルプロセスを使用しないアプローチを示す。このアプローチは、システム内のＮＩＣドライバ間の協同にのみ依存している。この機能をドライバに実装することは、開発コストを増加させるが、高速な応答を可能にするという利点がある。（ドライバは

、他のノードからの入力プロトコルメッセージに応答するのに、割り込みの処理を利用することができる。）更に、ドライバは、そのＲＤＭＡ機能を利用可能とする前に、処理途中である古いＲＤＭＡアクセスが無いことを確保することができる。要求されるロジックは、ユーザレベルプロセスによって要求されるものより、おそらくは簡単で、従って、信頼性も改善されるであろう。

－システム構成

図１は、本発明の実施形態が基礎とするシステム構成を示す図である。

【００１４】

システムは、多くのホスト１０－１～１０－７によって構成されている。各ホスト１０－１～１０－７は、１以上のＮＩＣ１１を持っている。ＮＩＣ１１は、ネットワーク１２を介して接続される。ＮＩＣ１１は、別々に、ＳＥＮＤ／ＲＥＣＥＩＶＥ機能１１－１とＲＤＭＡ機能１１－２の両方を持つ。例えば、ＳＥＮＤ／ＲＥＣＥＩＶＥ機能１１－１を利用可能とし、同時にＲＤＭＡ機能１１－２を利用不可能とすることが可能である。

【００１５】

ＮＩＣ１１は、専用のハードウェアのアドレス翻訳保護テーブル（ＴＰＴ）を持たない。そのかわり、ドライバ１３は、協同して、ソフトウェアのシステム全体に渡る保護テーブル１３－１を持つ。テーブル１３－１は、ローカル部とリモート部の２つの部分からなる。ローカルなユーザアプリケーションがドライバ１３にメモリを登録する結果として、テーブルのローカル部にエントリが加えられる。ユーザプロセスは、登録したメモリへのＲＤＭＡアクセスを利用可能としたり利用不可能とすることができる。ドライバ１３は、テーブル１３－１のローカル部の変更を他のノードのテーブル１３－１の対応するリモート部に反映させる。再起動するホストが無い限り、全てのドライバ１３は、保護テーブル１３－１に同じ情報を持っている。

【００１６】

各ドライバの保護テーブル１３－１は、システム内の全てのＮＩＣ１１についての情報を含んでいる。ローカルプロセスがＲＤＭＡ動作を発行すると、ドライバ１３は、保護テーブル１３－１のコピーをチェックし、リモートアクセスが許

可されているか否かを判断する。（保護テーブルを他の形で実現することも可能である。例えば、ドライバは、各ＲＤＭＡトランザクションを異なるプロトコルで有効化する。この場合、各ドライバは、ローカルアドレスのみを登録するテーブルを持ち、リモートアクセスが要求されたときには、リモートアクセスの要求が発行される前に、プロトコルを確立するメッセージが、ローカルアドレスのみを有するテーブルを使って、ホスト間でやりとりされる。しかし、古い保護情報を使用するという問題は依然存在する。以後に示すアプローチは依然適用可能である。）

システム内のＮＩＣ１１の最大数は、全てのホスト１０－１～１０－７に知られているか、確定可能である。この最大数は、現在接続されているＮＩＣ１１データはなく、ネットワーク１２に接続可能なＮＩＣ１１の可能な全数を反映している。メッセージあるいは、ＲＤＭＡ動作は、ＮＩＣ１１が適切に初期化されていないあるいは、ネットワーク１２に接続されていないのでない限りは、データを配布することになる。後者の場合、メッセージを送信したドライバ１３が見ることができる、ハードウェアで生成された、ネガティブアクノレッジメント（ＮＡＣＫ）がある。

【００１７】

ドライバ１１がＮＡＣＫを検出すると、ＮＩＣ１１がシステムに再組み込みされるまで、リモートＮＩＣ１１へのアクセスをドライバが禁止する。

最後に、同じ送信元、宛先のＲＤＭＡ及びＳＥＮＤ動作は、同じ経路を使用する。ネットワーク１２は、このような動作が互いに追い越すことを許さない。

【００１８】

図２は、ホストの構成を示す図である。

ホスト１０－ｉと１０－ｊは、ＮＩＣ１１を介して、ネットワーク１２に接続される。ホスト１０－ｉと１０－ｊは、ユーザ空間とカーネル空間を有している。プロセスＰは、ユーザ空間に有る。ドライバは、カーネル空間にある。ＮＩＣは、ホスト１０－ｉと１０－ｊのハードウェアの一部である。各ＮＩＣ１１は、ネットワーク１２とドライバ１３のインターフェースを持ったＳＥＮＤ／ＲＥＣＥＩＶＥ機能１１－１とＲＤＭＡ機能１１－２を有している。

【００１９】

プロセスＰが起動されると、プロセスＰは、ドライバ１３のＲＤＭＡ保護機能内の翻訳保護テーブル（ＴＰＴ）１３－１にプロセスＰ自身が使用するメモリ領域ｉを登録する。そして、ＴＰＴ１３－１は、プロセスＰに割り当てられた領域ｉのエントリを有する。同様に、ホスト１０－ｊにおいては、プロセスＭは、メモリ領域Ｘをドライバ（不図示）に登録する。この登録は、ホスト１０－ｉのＴＰＴ１３－１に反映される。そして、ホスト１０－ｉのＴＰＴは、プロセスＭの領域Ｘのエントリを持つようになる。

【００２０】

プロセスＰがプロセスＭにアクセスする必要があるときは、プロセスＰは、ＴＰＴ１３－１の領域Ｘのエントリを参照して、アクセスのための要求を発行し、アクセスが確立される。プロセスＭがクラッシュしたか、ホスト１０－ｊがクラッシュした場合には、ＴＰＴ１３－１の領域Ｘのエントリは、ホスト１０－ｊのプロセスＭが利用できないことを示すために、ＴＰＴ１３－１から消去される。

【００２１】

図３は、翻訳保護テーブルの構成を示す図である。

図３においては、ネットワーク内に、０～Ｎ－１まで番号付けられたＮ個のノードがあるとする。

【００２２】

ＴＰＴは、上記したように、ローカル部とリモート部を有している。ノード０では、ＴＰＴのローカル部に、自ノード０のためのメモリ領域が登録されている。リモート部には、ノード１～ノードＮ－１のためのメモリ領域が登録されている。ＴＰＴのエントリフォーマットが図３（ｂ）に示されている。ＴＰＴの各エントリは、項目として、論理アドレス、物理アドレス、長さ、保護属性を有している。ユーザ空間のユーザプロセスは、論理アドレスを用いて、アクセス要求を発行する。この論理アドレスは、ＴＰＴを参照することによって、物理アドレスに変換され、物理アドレスを用いて、宛先にメッセージを送信する。長さの項目は、物理アドレスの項目で特定されるアドレスの領域の長さを、例えば、バイト単位で示すものである。保護属性の項目は、物理アドレスで示される領域の、メ

モリ保護の種類（読み取り専用、書き込み専用など）を特定するものである。

【００２３】

図４は、本発明の実施形態に従ったメッセージ交換のシーケンスを示す図である。

ホストがドライバ（ＤＲＩＶＥＲ＿Ａ）を起動したとき、初期化の一部として、ＳＥＮＤ／ＲＥＣＥＩＶＥ機構を利用可能とし、ＲＤＭＡ機構を利用不可能とする。ＳＥＮＤ／ＲＥＣＥＩＶＥ機構は、ＮＩＣドライバがリモートノードと通信できるように利用可能とされる。ＲＤＭＡ機構は、あらゆる入力する（古い）ＲＤＭＡ動作を拒絶するために、利用不可能とされる。

【００２４】

ＤＲＩＶＥＲ＿Ａは、システム全体に、起動していることを示すメッセージ（ＢＯＯＴＩＮＧ）をブロードキャスト（あるいは、ユニキャスト）する。他のＮＩＣドライバから、その存在を認識したというメッセージを得るまでまつ。各ホストの応答が一回のみカウントされるように、応答を監視する。

【００２５】

ＤＲＩＶＥＲ＿ＡからのＢＯＯＴＩＮＧメッセージを受信したＮＩＣドライバ（ＤＲＩＶＥＲ＿Ｂ）は、ローカルに保持している翻訳保護テーブルのＤＲＩＶＥＲ＿Ａの部分をクリアする。テーブルのＤＲＩＶＥＲ＿Ａの部分をクリアすることは、ＤＲＩＶＥＲ＿ＡのＮＩＣに対するＲＤＭＡアクセスを禁止する効果がある。また、ＤＲＩＶＥＲ＿Ａの翻訳保護テーブルへの更新を送信することも禁止する。（ＤＲＩＶＥＲ＿Ｂが、より早くに、ＤＲＩＶＥＲ＿ＡのＮＩＣが機能しておらず、既にそれへのアクセスが禁止されていることを認識することは可能である。アクセスが既に禁止されているか否かは、ここでは問題ではない。）次のステップでは、ＤＲＩＶＥＲ＿Ｂが、そのＮＩＣへの全ての発行された動作、ＳＥＮＤあるいはＲＤＭＡが実行されたか否かを確認する。その後、ＤＲＩＶＥＲ＿Ｂは、アクノレッジメントメッセージ（ＡＣＫ）で応答する。

【００２６】

ＤＲＩＶＥＲ＿Ａは、ＤＲＩＶＥＲ＿ＢにＢＯＯＴＩＮＧメッセージを送信し、ＤＲＩＶＥＲ＿ＢがＡＣＫメッセージで応答したので、ＤＲＩＶＥＲ＿ＡとＤ

ＲＩＶＥＲ＿Ｂとの間に、まだ終わっていない、あるいは、実行中のＲＤＭＡ動作が無いことを保証する。（追い越しなし、及び、同じ経路を使っていることを仮定。）換言すれば、ＤＲＩＶＥＲ＿ＡのＮＩＣは、古いＲＤＭＡ動作から保護されている。ＤＲＩＶＥＲ＿Ｂが、ＡＣＫを送る前に、全ての発行された動作が実行されたことを確認しないならば、この保証は成り立たない。ＡＣＫが、以前に発行されたＲＤＭＡ動作の前に送られることはありえる。特に、ＮＩＣが発行された動作を処理するのに、複数のキューを持っている場合、これがあり得る。

【００２７】

リモートＮＩＣが接続されていない、あるいは、動作していない場合、ドライバのＡＣＫではなく、ハードウェアで生成された（ＨＷ）ＮＡＣＫが受信される。従って、ドライバは、システム内の全てのノードを数えることができ、どのＮＩＣがシステムの一部で、どれがそうでないかを知ることができる。全てのＮＩＣが数えられた時にのみ、ＤＲＩＶＥＲ＿Ａは、ＲＤＭＡ機能を利用可能にする。そして、ＲＤＭＡアクセスの準備ができたことを示すメッセージ（ＲＤＭＡ＿ＲＥＡＤＹ）を分かっている全てのＮＩＣに送信する。

【００２８】

ＲＤＭＡ＿ＲＥＡＤＹメッセージを受信するＮＩＣドライバ（ＤＲＩＶＥＲ＿Ｂ）は、そのＮＩＣへのＲＤＭＡアクセスを利用可能にする。言い換えれば、ＤＲＩＶＥＲ＿Ａのシステム保護テーブルに、翻訳保護テーブルの更新を送信し始める。そして、ＤＲＩＶＥＲ＿Ａの保護テーブルの内、ＤＲＩＶＥＲ＿ＢのＮＩＣを参照する部分を更新する。全てのリモートＮＩＣドライバが、ＤＲＩＶＥＲ＿Ａのテーブルのこれらの部分を更新したとき、ＤＲＩＶＥＲ＿Ａは、システムに完全に再組み込みされる。

【００２９】

ホストの状態は、何時変わるかわからないので、ドライバがそのホストのクラッシュに応答する前に、ＮＩＣドライバ（ＤＲＩＶＥＲ＿Ｃ）によってＢＯＯＴＩＮＧメッセージが受信されることもあり得る。この場合、ＤＲＩＶＥＲ＿Ａは、ＤＲＩＶＥＲ＿ＣのＮＩＣを数えることができない。従って、ＤＲＩＶＥＲ＿Ａは、適切なタイムアウト時間後、数えることができなかったＮＩＣにＢＯＯＴ

ＩＮＧメッセージを再送する。この処理は、全てのＮＩＣが数えられるまで繰り返される。別の方法として、ＤＲＩＶＥＲ＿ＣからＡＣＫメッセージの代わりに、ＢＯＯＴＩＮＧメッセージを受け取った場合には、ＤＲＩＶＥＲ＿Ａは、このプロセスをやらないことも可能である。

【００３０】

新しいホストは、再起動したホストと同様の方法でシステムに加えることが可能である。

図５は、本発明の実施形態に従った、ＴＰＴにユーザプロセスが使う領域をユーザプロセスが登録する処理のフローチャートである。

【００３１】

ステップＳ１においては、ノードＮのユーザプロセスは、ドライバに登録されるべき領域の情報を与える。ステップＳ２においては、ノードＮのドライバは、領域を特定し（あるいは、割当て）、論理アドレスを物理アドレスに変換し、ローカルＴＰＴにエントリｉを割当て、その領域に関する情報を格納する。ステップＳ３においては、ノードＮのドライバは、ネットワークにつながれた全てのノードに更新要求を発行する。更新要求により、ノードＮのドライバによって割り当てられた領域に対応する各ノードのＴＰＴのエントリｉを更新する。

【００３２】

図６は、本発明の実施形態に従った、ユーザプロセスがＲＤＭＡ　Ｗｒｉｔｅ転送コマンドを発行する処理のフローチャートである。

ステップＳ１０において、ユーザプロセスは、ローカルな登録された領域Ｘ（以下、ＲＲ＿Ｘ）から、遠隔の登録された領域Ｙ（以下、ＲＲ＿Ｙ）へＢバイトデータ（Ｂは、バイト単位のデータ長である）を転送する要求である、ＲＤＭＡコマンドをドライバへ発行する。ステップＳ１１においては、ドライバは、ユーザプロセスが、ＲｅａｄアクセスをＲＲ＿Ｘに発行したか否かを判断する。ステップＳ１１の判断がＮＯの場合には、ドライバは、ユーザプロセスに保護エラーを返す。ステップＳ１１の判断がＹＥＳの場合には、ステップＳ１２に進む。ステップＳ１２においては、ドライバは、ユーザプロセスがＲＲ＿Ｙに対し、Ｗｒｉｔｅアクセスを発行したか否かを判断する。ステップＳ１２の判断がＮＯの場

合には、ドライバは、ユーザプロセスに保護エラーを返す。ステップＳ１２の判断がＹＥＳの場合には、ステップＳ１３に進む。ステップＳ１３においては、ドライバは、転送されたデータの長さが、２つの領域（すなわち、ＲＲ＿ＸとＲＲ＿Ｙ）の限界以内であるか否かを判断する。ステップＳ１３の判断がＮＯの場合には、ドライバは、ユーザプロセスに保護エラーを返す。ステップＳ１３の判断がＹＥＳの場合には、ドライバは、ＮＩＣにデータを転送させるため、ＮＩＣに対し、転送コマンドを発行する。

【００３３】

図７は、本発明の実施形態に従った、ホストの起動処理を示すフローチャートである。

ホストが起動された後、ドライバは、ＮＩＣとＴＰＴを初期化し、ＳＥＮＤ／ＲＥＣＥＩＶＥ機能を利用可能とし、ＲＤＭＡ機能を利用不可能とし、｛応答集合｝という変数を全ての０に設定する（ステップＳ１５）。ステップＳ１６においては、ドライバは、全てのノードにＢＯＯＴＩＮＧメッセージを送信する。ステップＳ１７では、ドライバは、所定の時間内で、他のノードからの応答を待つ。ステップＳ１７でタイムアウトが起こったときは、ステップＳ１６に戻る。ステップＳ１７において、ドライバが、応答メッセージ（ＡＣＫ、ハードウェアで生成されたＮＡＣＫ、ＢＯＯＴＩＮＧメッセージのいずれか）を受信した場合には、取り阿波は、応答メッセージの発信ノードを特定し、ステップＳ１８において、｛応答集合｝にノードの識別子を加える。

【００３４】

例えば、応答集合は、０と１の列から構成される。ステップＳ１５においては、この列は全て０に設定される。この列中、各位は、ホストに対応付けられる。自ホストのドライバによって応答が受信された場合には、応答を送信したホストが特定され、対応する位置にある列内の０を１に設定する。列が１で満たされた場合、これは、全てのホストから応答を受信したことを示す。

【００３５】

ステップＳ１９において、ドライバは、全てのノードが｛応答集合｝にあるか否かを判断する。ステップＳ１９における判断が、ＮＯの場合には、ステップＳ

１７に進む。ステップＳ１９の判断がＹＥＳの場合には、ステップＳ２０に進む。ステップＳ２０では、ドライバは、ＲＤＭＡ機能を利用可能とし、ステップＳ２１において、全てのノードにＲＤＭＡ＿ＲＥＡＤＹメッセージを送信する。

【００３６】

図８は、起動されたノード以外のノードの動作を示すフローチャートである。

ステップＳ２５において、ドライバは、起動されたノードＮのドライバからＢＯＯＴＩＮＧメッセージを受信する。ステップＳ２６において、ドライバはノードＮのＴＰＴエントリをクリアし、ノードＮへのＴＰＴ更新を利用不可能にする。ステップＳ２７においては、ドライバは、ノードＮに対し、ＡＣＫで応答する。他の場合には、ノードＮ以外のノードのドライバは、ダウンしている可能性がある。この場合、ＢＯＯＴＩＮＧメッセージへの応答（ＨＷ　ＮＡＣＫ）は、ハードウェアで生成され、、ノードＮに送り返される。また、ノードＮ以外のノードは、ちょうど起動されたばかりかもしれない。この場合、ＢＯＯＴＩＮＧメッセージがノードＮに送られ、このノードがノードＮとして動作する。

【００３７】

いずれにしても、ドライバは、ステップＳ２８において、ＲＤＭＡ＿ＲＥＡＤＹメッセージをノードＮから受信する。ステップＳ２９においては、ドライバは、ノードＮに対するＴＰＴ更新を利用可能とする。そして、ステップＳ３０において、ドライバは、そのローカルＴＰＴ部の内容により、ノードＮのＴＰＴを更新する。この更新は、例えば、ＲＤＭＡ　Ｗｒｉｔｅコマンドにより達成される。

【００３８】

処理フローは、２つのホストのみを参照して説明したが、この処理フローは、多くのホスト間で実行される。

（付記１）ＲＤＭＡ機能を有した、ホストと複数のホストからなるネットワーク内の該ホストに設けられた装置であって、

該ネットワーク内のホストが起動したとき、該ホストが起動したことを示す第１のメッセージを、該ネットワーク内の全ての該複数のホストに送信する手段と、

該複数のホストからの該ホストへのＲＤＭＡアクセスを利用不可能とする手段と、

該ホストに第２のメッセージを送信することによって、第１のメッセージに応答する手段と、

該ホストが、該複数のホストからのＲＤＭＡアクセスを受付可能であることを示す第３のメッセージを、該複数のホストの全てから第２のメッセージを受信し、ＲＤＭＡ機能を利用可能とした後、該複数のホストの全てに送信する手段と、を備えることを特徴とする装置。

【００３９】

（付記２）前記装置は、前記ホストのドライバに含まれることを特徴とする付記１に記載の装置。

（付記３）他のホストにＲＤＭＡアクセスを行うための情報を有する翻訳保護テーブル手段を更に有し、

前記第１のメッセージを受信したとき、該第１のメッセージを送信したホストに関する情報を該翻訳保護テーブル手段からクリアし、該ホストへのＲＤＭＡアクセスを不可能にすることを特徴とする付記１に記載の装置。

【００４０】

（付記４）前記複数のホストの翻訳保護テーブルは、該複数のホストへ、前記第３のメッセージが送られた後、更新されることを特徴とする付記３に記載の装置。

【００４１】

（付記５）前記第２のメッセージは、アクノレッジメント、ノンアクノレッジメント、及び、前記複数のホストから送られた前記第１のメッセージの一つであり、ノンアクノレッジメントは、ハードウェアによって生成されることを特徴とする付記１に記載の装置。

【００４２】

（付記６）前記第２のメッセージが前記複数のホストの全てから受信されたか否かは、監視され、０と１の列からなる応答集合によって判断されることを特徴とする付記１に記載の装置。

【００４３】

（付記７）前記ホストは、ＲＤＭＡ機能と他のメッセージ通信機能を有するネットワークインターフェースカードを有し、ＲＤＭＡ機能と他のメッセージ通信機能の初期化は独立して行われることを特徴とする付記１に記載の装置。

【００４４】

（付記８）ＲＤＭＡ機能を有した、ホストと複数のホストからなるネットワーク内の該ホストの方法であって、

該ネットワーク内のホストが起動したとき、該ホストが起動したことを示す第１のメッセージを、該ネットワーク内の全ての該複数のホストに送信するステップと、

該複数のホストからの該ホストへのＲＤＭＡアクセスを利用不可能とするステップと、

該ホストに第２のメッセージを送信することによって、第１のメッセージに応答するステップと、

該ホストが、該複数のホストからのＲＤＭＡアクセスを受付可能であることを示す第３のメッセージを、該複数のホストの全てから第２のメッセージを受信し、ＲＤＭＡ機能を利用可能とした後、該複数のホストの全てに送信するステップと、

を備えることを特徴とする方法。

【００４５】

（付記９）前記方法は、前記ホストのドライバで行われることを特徴とする付記８に記載の方法。

（付記１０）他のホストにＲＤＭＡアクセスを行うための情報を格納する翻訳保護テーブルステップを更に有し、

前記第１のメッセージを受信したとき、該第１のメッセージを送信したホストに関する情報を該翻訳保護テーブルステップで格納された情報からクリアし、該ホストへのＲＤＭＡアクセスを不可能にすることを特徴とする付記８に記載の方法。

【００４６】

（付記１１）前記複数のホストの翻訳保護テーブルは、該複数のホストへ、前記第３のメッセージが送られた後、更新されることを特徴とする付記１０に記載の方法。

【００４７】

（付記１２）前記第２のメッセージは、アクノレッジメント、ノンアクノレッジメント、及び、前記複数のホストから送られた前記第１のメッセージの一つであり、ノンアクノレッジメントは、ハードウェアによって生成されることを特徴とする付記８に記載の方法。

【００４８】

（付記１３）前記第２のメッセージが前記複数のホストの全てから受信されたか否かは、監視され、０と１の列からなる応答集合によって判断されることを特徴とする付記８に記載の方法。

【００４９】

（付記１４）前記ホストは、ＲＤＭＡ機能と他のメッセージ通信機能を有するネットワークインターフェースカードを有し、ＲＤＭＡ機能と他のメッセージ通信機能の初期化は独立して行われることを特徴とする付記８に記載の方法。

【００５０】

【発明の効果】

本発明によれば、ＲＤＭＡ機能を有するＮＩＣを持つホストを、ユーザレベルプロセスを用いずに、ネットワークに安全に組み込む装置が提供される。

【図面の簡単な説明】

【図１】

本発明の実施形態が基本とするシステム構成を示す図である。

【図２】

ホストの構成を示す図である。

【図３】

翻訳保護テーブル（translation and protection table）の構成を示す図である。

【図４】

本発明の実施形態に従った、メッセージ交換のシーケンスを示す図である。

【図５】

本発明の実施形態に従った、ユーザプロセスが、ユーザプロセスが使用する領域をＴＰＴに登録する処理のフローチャートである。

【図６】

本発明の実施形態に従った、ユーザプロセスがＲＤＭＡ転送コマンドを発行する処理のフローチャートである。

【図７】

本発明の実施形態に従った、ホストの起動処理のフローチャートである。

【図８】

起動されたノード以外のノードの動作を示すフローチャートである。

【符号の説明】

１０－１～１０－７、１０－ｉ、１０－ｊ　　ホスト

１１　　ＮＩＣ

１１－１　　ＳＥＮＤ／ＲＥＣＥＩＶＥ機能

１１－２　　ＲＤＭＡ機能

１２　　ネットワーク

１３　　ドライバ

１３－１　　ＲＤＭＡ保護機能

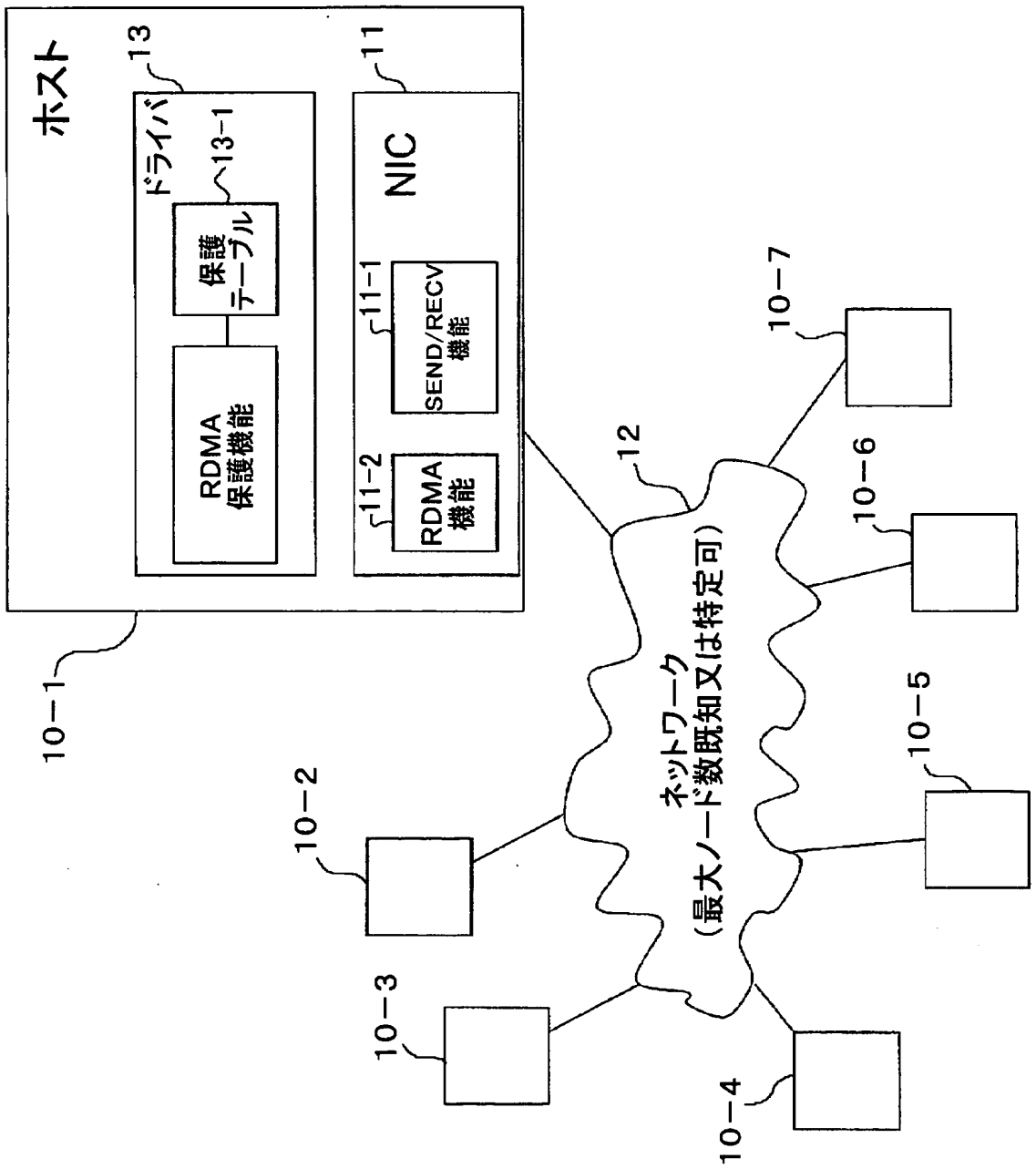
２０　　ユーザ空間

２１　　カーネル空間

【書類名】　　図面

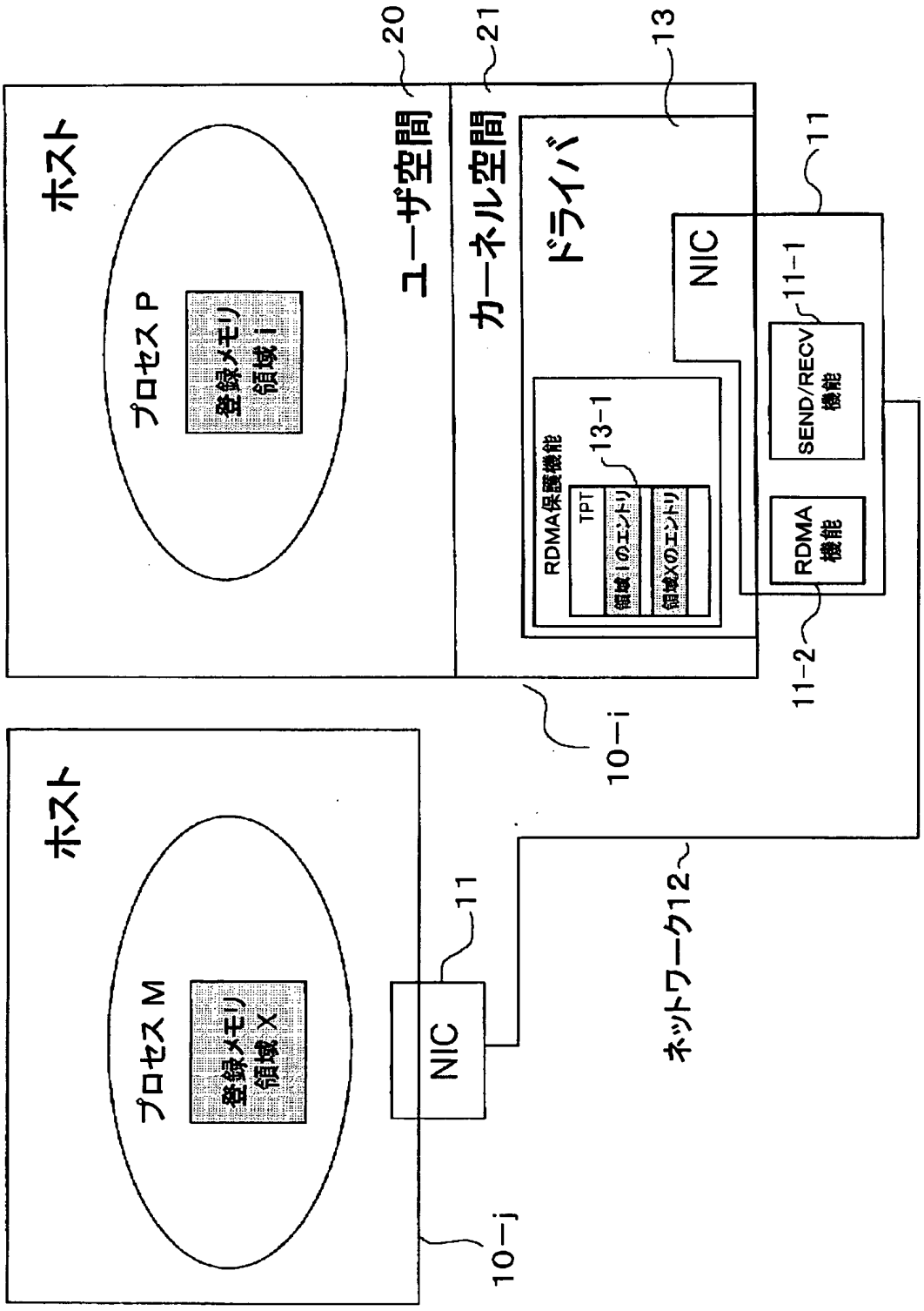
【図１】

本発明の実施形態が基本とするシステム構成を示す図



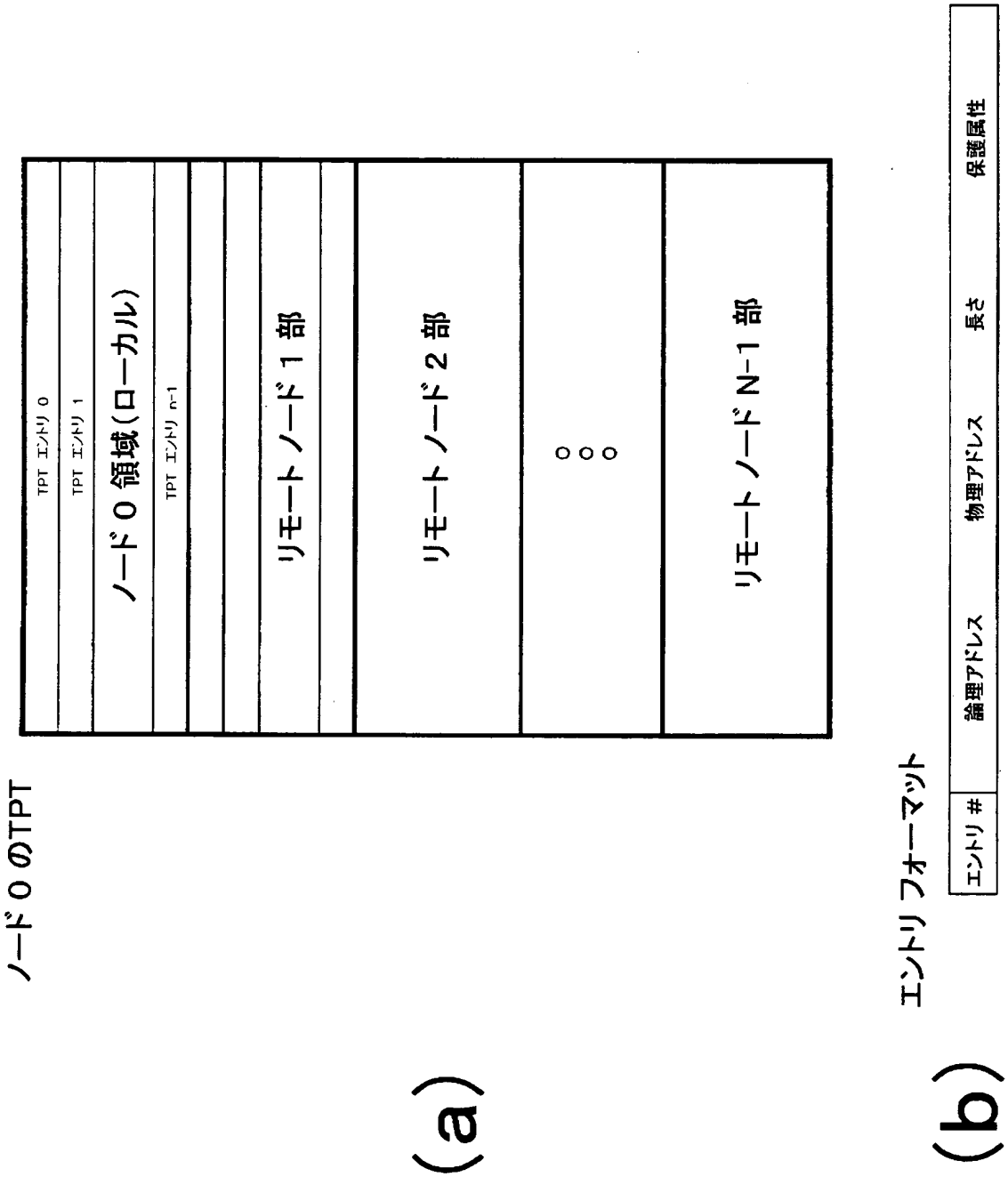
【図２】

ホストの構成を示す図



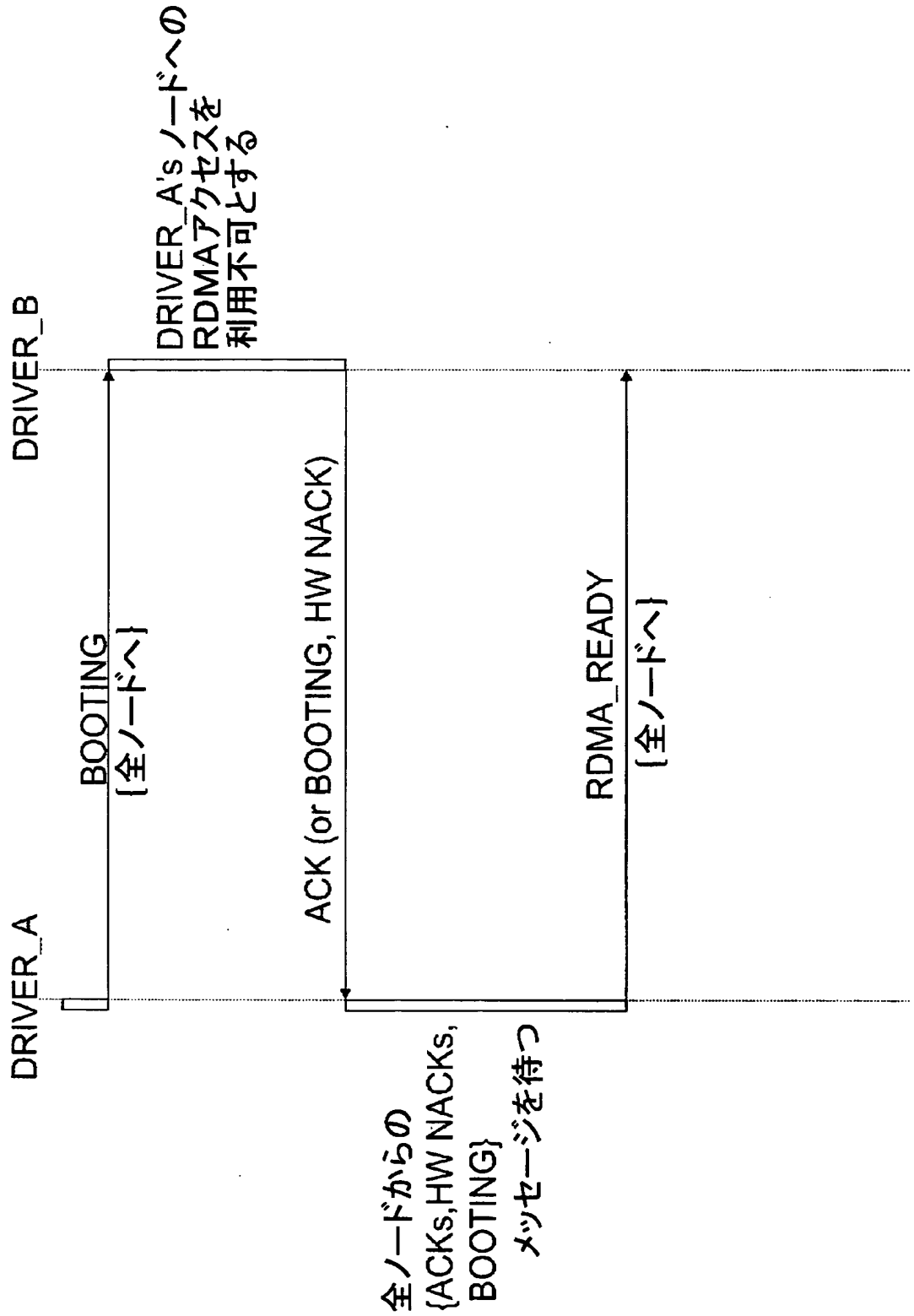
【図３】

翻訳保護テーブル（translation and protection table）の構成を示す図



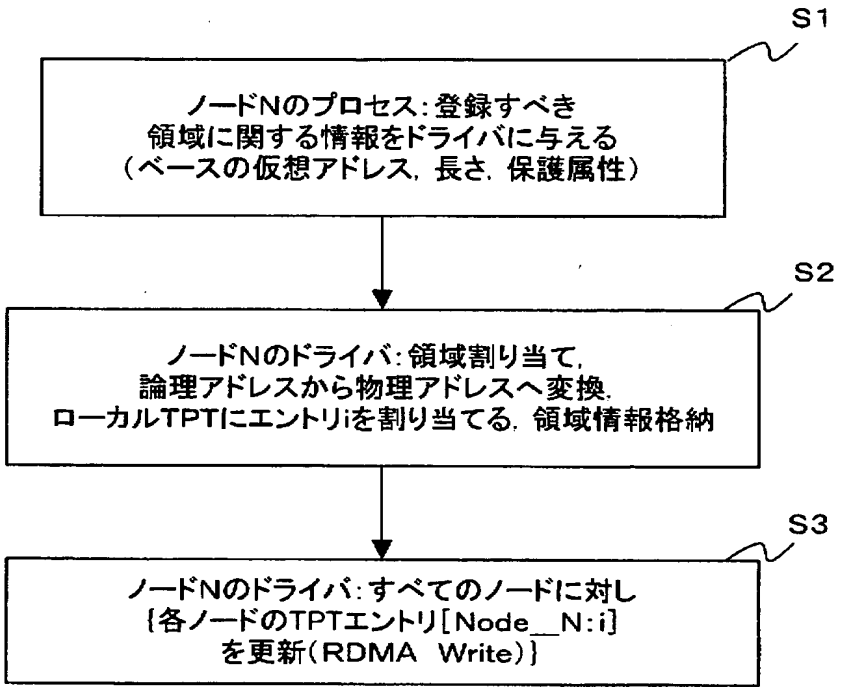
【図４】

本発明の実施形態に従った、
メッセージ交換のシーケンスを示す図



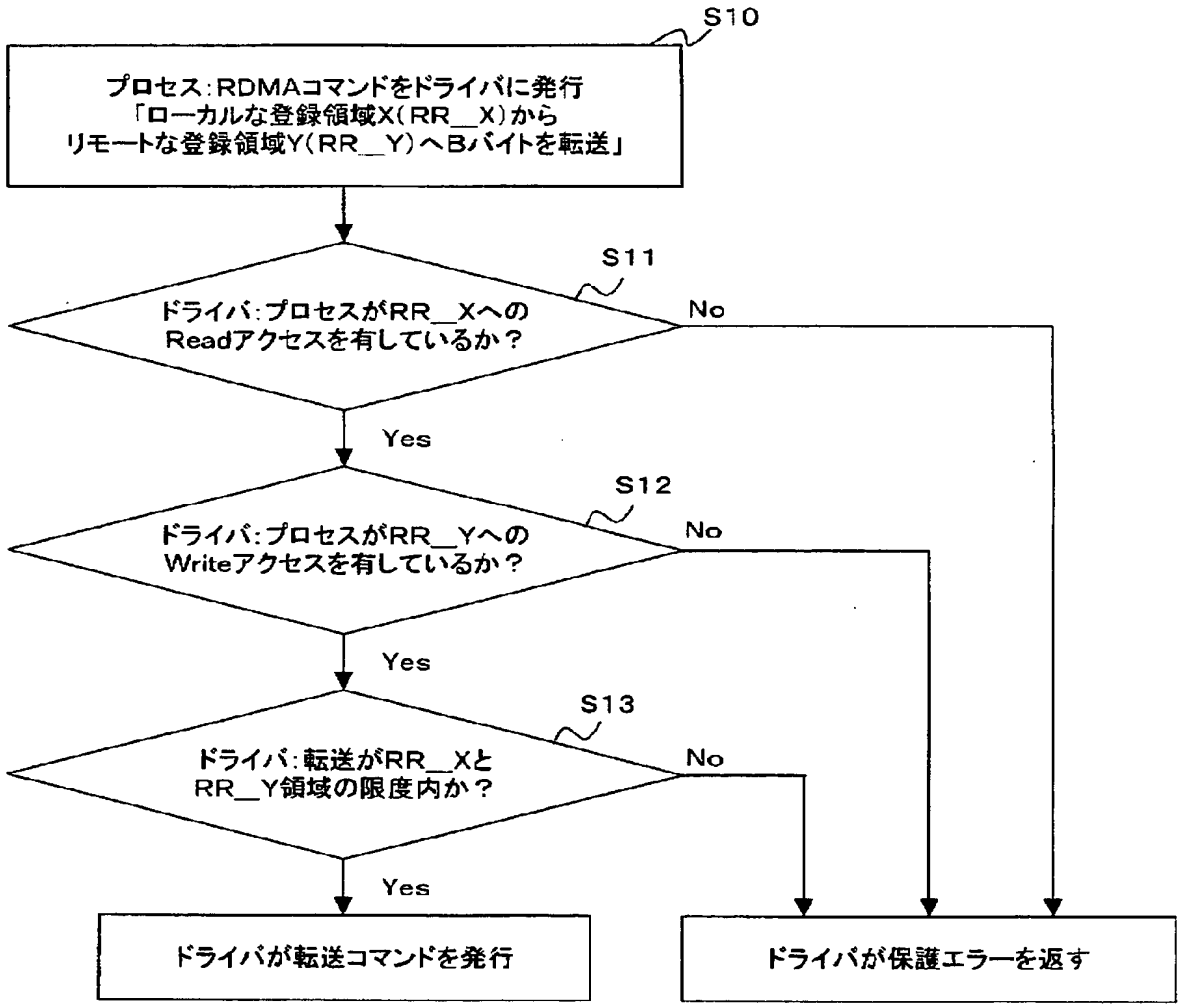
【図５】

本発明の実施形態に従った、ユーザプロセスが、
ユーザプロセスが使用する領域を
TPTに登録する処理のフローチャート



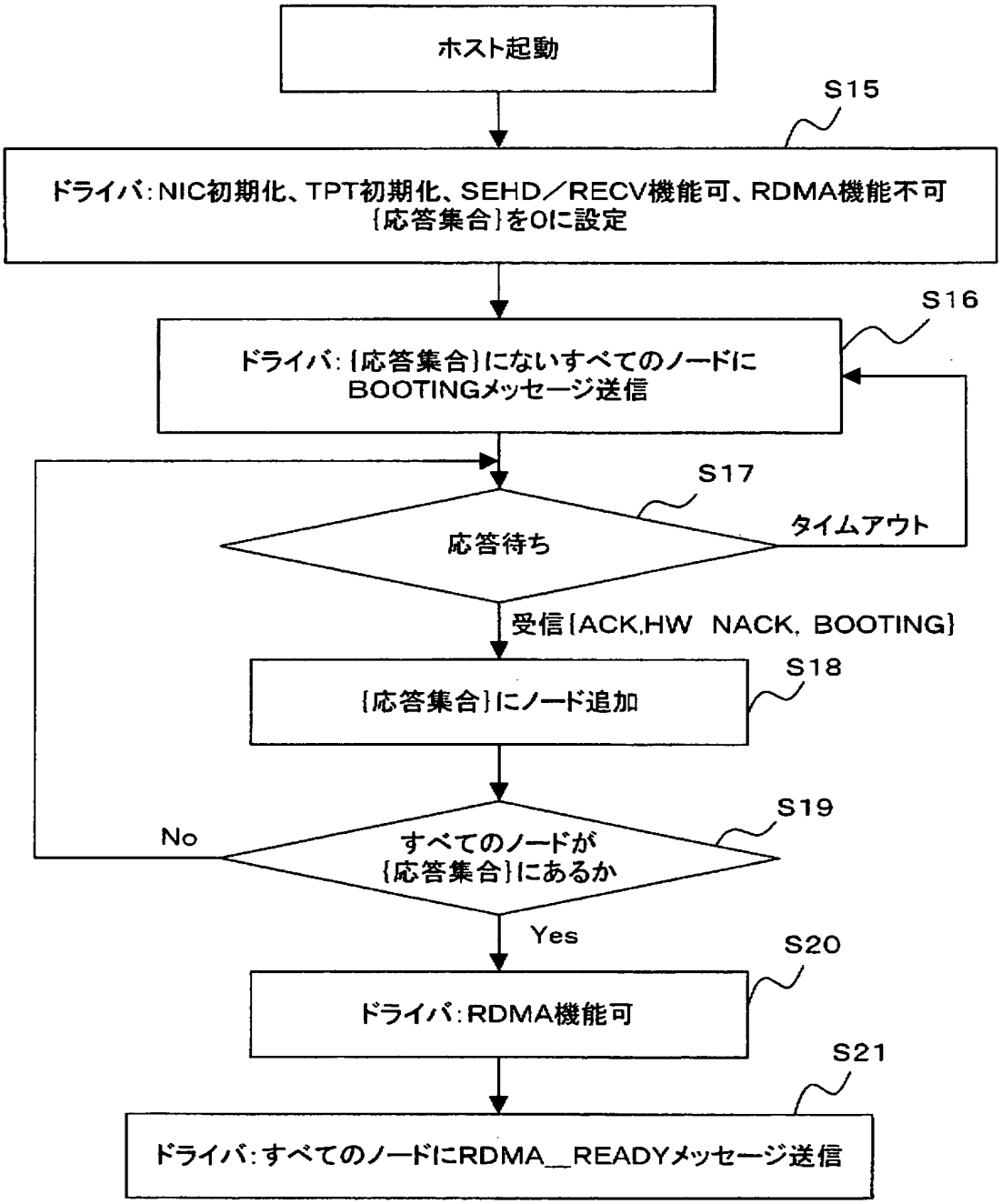
【図６】

本発明の実施形態に従った、ユーザプロセスが
RDMA転送コマンドを発行する処理のフローチャート



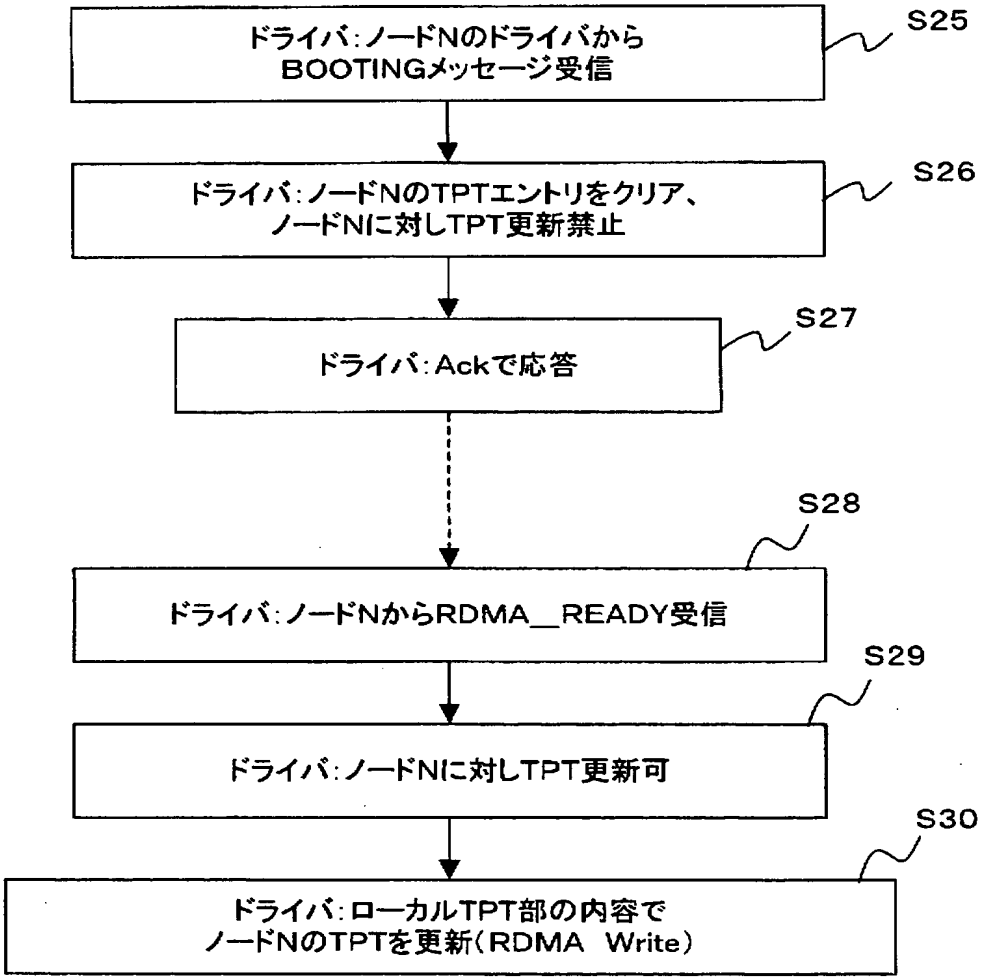
【図７】

本発明の実施形態に従った、
ホストの起動処理のフローチャート



【図８】

起動されたノード以外の
ノードの動作を示すフローチャート



【書類名】　　要約書

【要約】

【課題】ＲＤＭＡ機能を有するＮＩＣを持つホストをネットワークに安全に組み込むための装置を提供する。

【解決手段】複数のホストを含むネットワークにおいて、ホストが起動すると、該ホストのドライバは、該ホストが起動することを示すＢＯＯＴＩＮＧメッセージを全ての他のホストに送り、これを受信した、起動するホスト以外のホストのドライバは、起動するホストへのＲＤＭＡアクセスを利用不可能にする。起動するホスト以外のホストのドライバは、ＡＣＫあるいはＢＯＯＴＩＮＧメッセージで応答する。ある場合には、ネットワークがハードウェアＮＡＣＫを生成する。起動するホストが全てのホストから応答を受信すると、起動するホストのドライバは自ホストへのＲＤＭＡアクセスを許可する。

【選択図】　　図４

特願２００２－３５７４４９

出　願　人　履　歴　情　報

識別番号　　［０００００５２２３］

１．変更年月日　　１９９６年　３月２６日
［変更理由］　　住所変更
住　所　　神奈川県川崎市中原区上小田中４丁目１番１号
氏　名　　富士通株式会社